Econ 311: Behavioral and Experimental Economics

Prof. Jeffrey Naecker

Wesleyan University

Behavioral Economics and The Internet

Motivation

- The internet (and technology more generally) has greatly expanded the options for empirical economics
- Much more data being collected for empirical studies
  - 6,000 tweets per second
  - 41,000 Facebook posts per second
  - Terabytes of publicly available financial data every day
- Also many more platforms for running experiments
  - Social media companies running experiments essentially constantly
  - Lower barrier to entry for researchers though Amazon Mechanical Turk

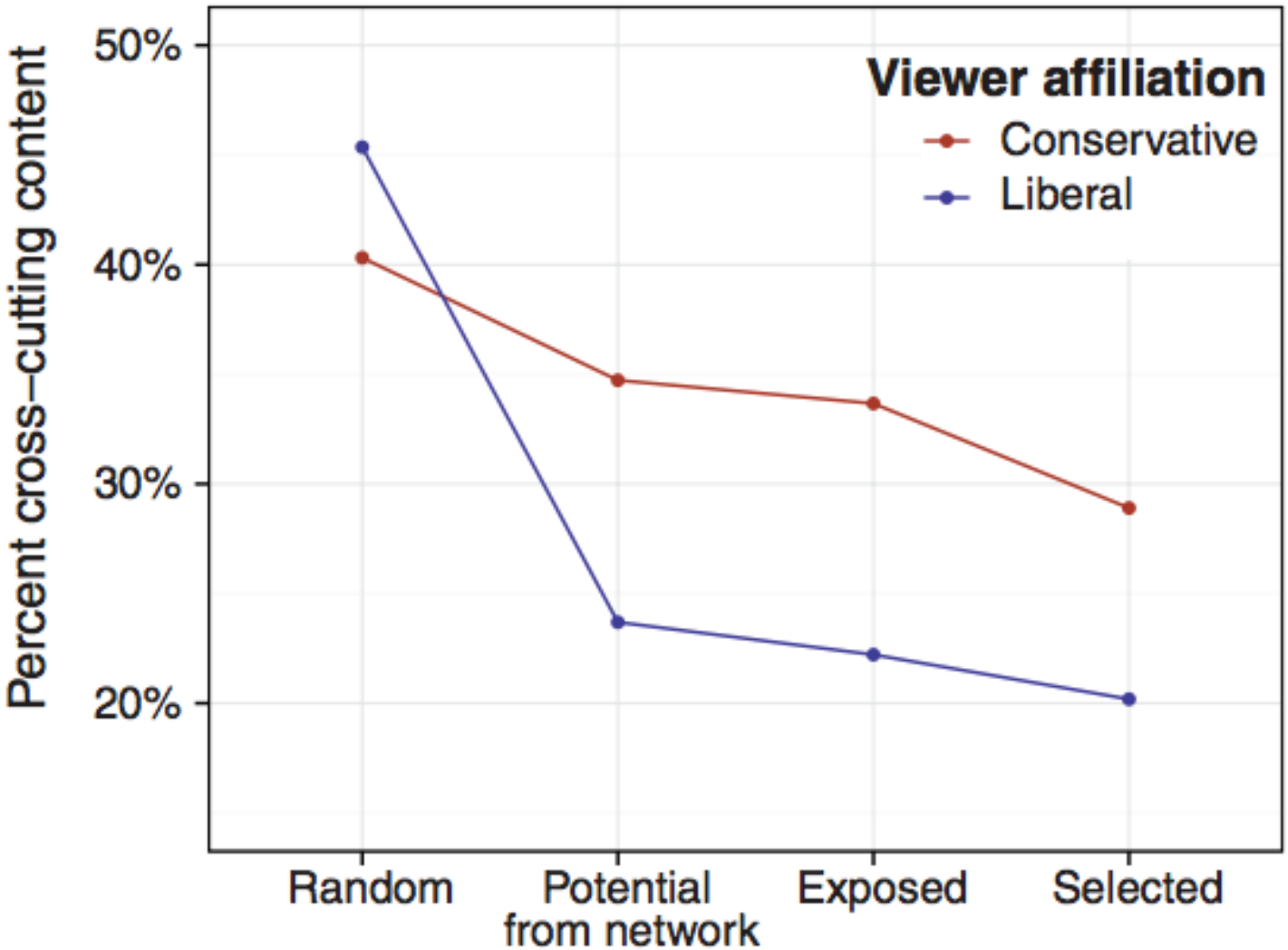
Is All This Useful?

- Question: does the internet make people better-informed?
  - Maybe yes:
    - Information is easier to obtain and verify
    - More likely to have conversations with people very different from yourself
  - Maybe not:
    - People may choose to surround themselves with connections and information sources that fit with their preferences
    - This is know as the *echo chamber effect*

Facebook Echo Chamber Study

- Bakshy, Messing, Adamic (2015) address this issue using data from Facebook posts
- Observed approx. 10 million people on Facebook (no experimental variation)
- Linked stories were classified either "cross-cutting" or "ideologically consistent" with each person's self-reported political affiliation
- What determines which content people read?
  1. Your network of friends
  2. How Facebook shows you your friends' content (Newsfeed)
  3. What content you choose to click on
- Baseline: how much cross-cutting content you would see if you were show random Facebook posts

Results from Adamic et al



Results from Adamic et al

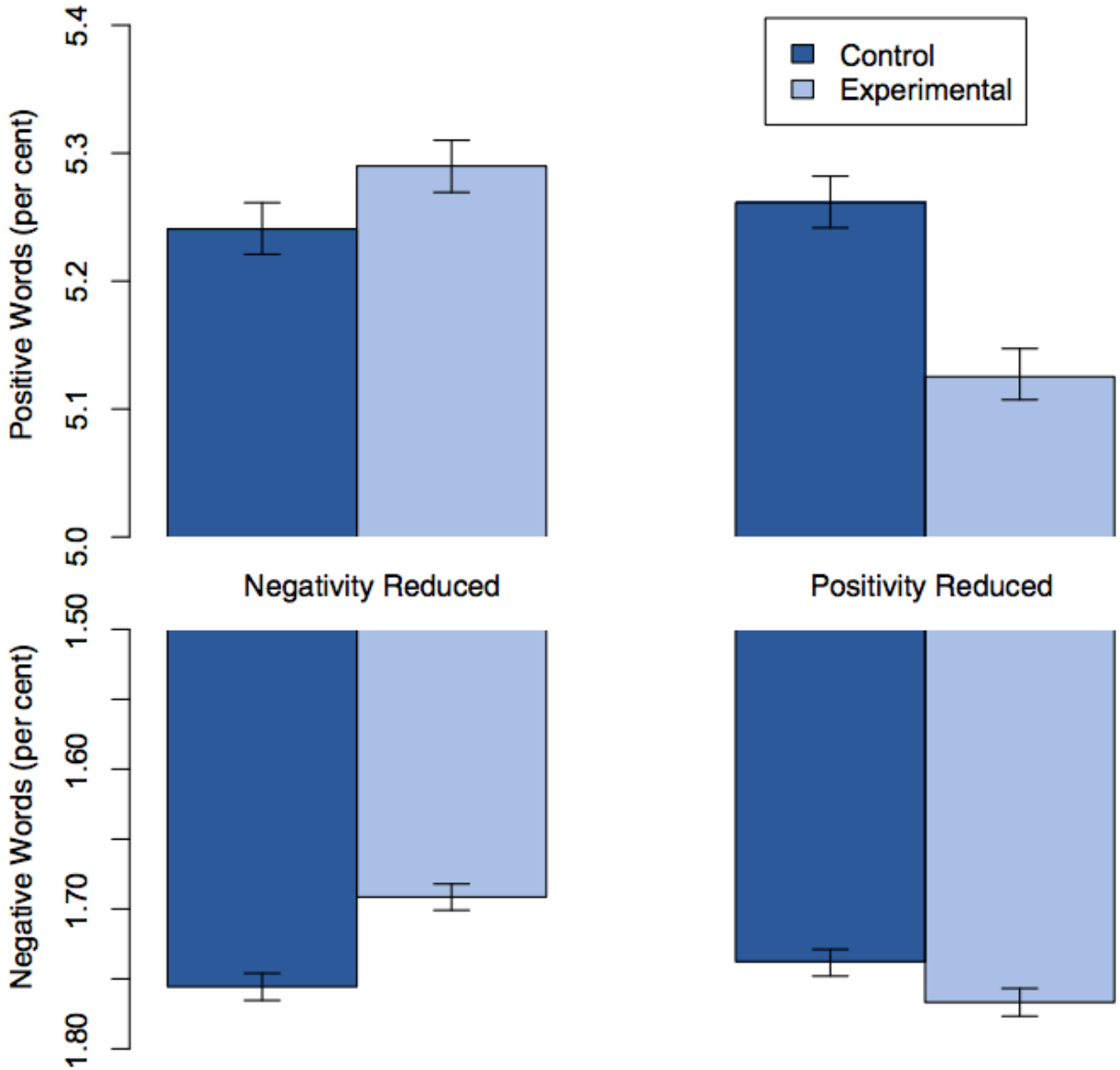
- Choice of friends is single biggest factor limiting exposure to cross-cutting content
  - This is the drop from "Random" to "Potential from Network"
- News feed algorithm has little effect on available content
  - This is the drop from "Potential from Network" to "Exposed"
- Selection from available content accounts for larger relative effect than algorithm
  - This is the drop from "Exposed" to "Selected" (ie clicked on)

Viewer affiliation	Random → Potential	Potential → Exposed	Exposed → Selected
Liberal	-0.626	-0.080	-0.063
Conservative	-0.212	-0.046	-0.172

News Feed Experiment

- The previous study used Facebook data but did not experimentally vary the user's experience
- Kramer, Guillory, and Hancock (2014) run experiment to determine how much of an effect news feed content has on user's emotions
- Experimental design:
  - Facebook posts categorized as either positive or negative
    - 22.4% negative, 46.8% positive
  - Treatment 1: Omit a percentage of all positive posts by friends that would otherwise show up on Newsfeed
  - Treatment 2: Omit a percentage of all negative posts by friends that would otherwise show up on Newsfeed
  - Controls: Omit a percentage of all posts
- Outcome variable: Positive/negative content of subjects' posts
- $N = 689,003$ people

Kramer et al Results



Kramer et al Results

- Results show emotional "contagion"
  - Omitting positive posts in feed lead to a 0.1% decrease in positive posts by subjects and a 0.04% increase in negative posts
  - Omitting negative posts in feed lead to a 0.07% decrease in negative posts by subjects and a 0.06% increase in positive posts
  - Results are statistically significant (due to large sample) but effect size is small
- Some public reaction to the paper was very negative, however:
  - One user on Twitter: "I wonder if Facebook KILLED anyone with their emotion manipulation stunt"
- What are some responses to these objections?
  - Note that Facebook gathered consent through terms of use agreement
  - No claim that the baseline algorithm is good or bad for mental health
  - One could argue that Facebook has an obligation to test their algorithm

Methodology: Amazon Mechanical Turk

- Most researchers do not have access to Facebook data (and certainly not able to manipulate their software)
- However, other tools do exist to reach lots of people online
- One such tool: Amazon Mechanical Turk
  - Online labor platform of English-speaking workers
  - Employers posts small tasks with an associated wage rate
  - Tasks can include experiments (either explicitly or implicitly)
  - Much cheaper and faster than running lab or field experiment
- Another tool: Harvard Digital Lab for the Social Sciences

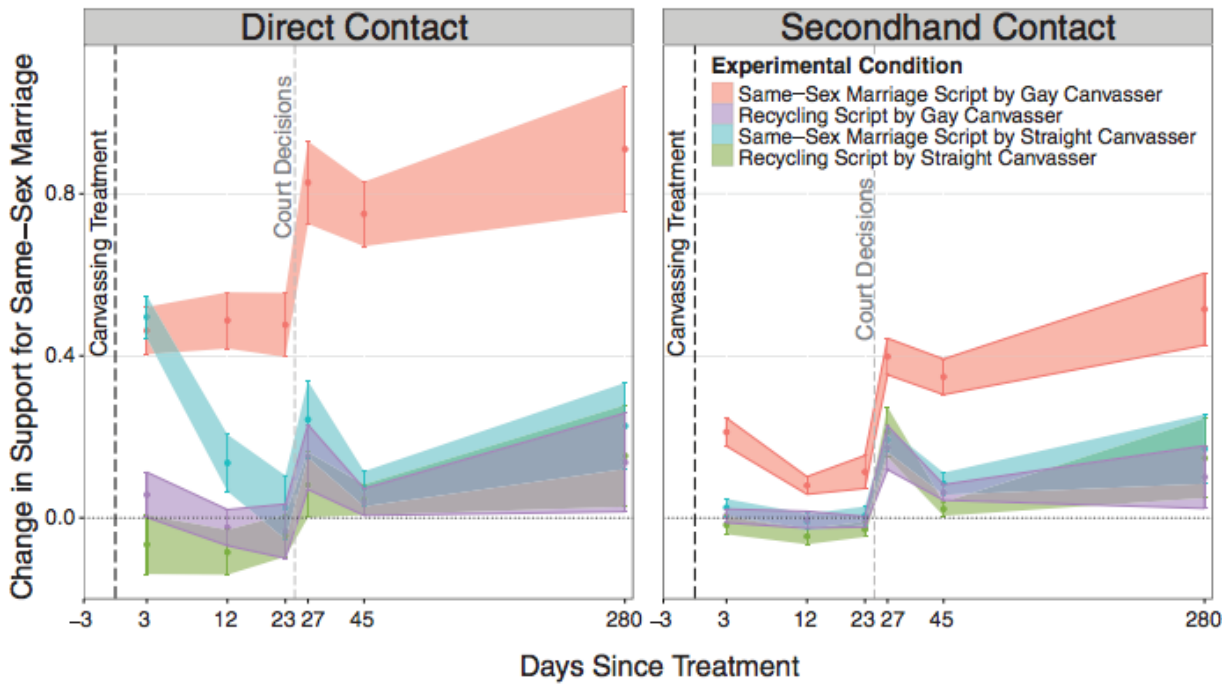
Reproducibility and Research Integrity

LaCour and Green (2014)

- We saw that the "echo chamber effect" can make it difficult for people's opinions to change?
- But forcing "cross-cutting" interactions might sway opinions
- La Cour and Green (2014) report an experiment attempting to change opinions on gay rights via canvasing
  - Initial baseline survey of opinions of voters in Los Angeles
  - Send either gay or straight canvaser to discuss gay rights with each voter for 22 minutes on average
  - Measure opinions on gay rights again with delay of 3 weeks, 5 weeks, and 9 months
  - Also measure opinions of people in the same household who did not talk directly to canvaser
  - Outcome: response on scale of 1-100, where 1=very cold and 100=very warm to idea of gay rights (thermometer scale)

Reported Results

- Both gay and straight canvassers were able to increase support for same-sex marriage
- Effect from gay canvassers persisted (or even increased) over time
- Gay canvassers also had an effect on other members in household



Just One Problem

- All the results reported by LaCour and Green (2014) were likely fabricated
- The deception appears to have been perpetrated entirely by LaCour (a graduate student at the time)
  - Canvassing was actually carried out as described by a non-profit (at great expense of time and money)
  - However, pre- and post-canvasing responses (allegedly collected via online surveys send to the canvassed households) were entirely made up by LaCour
  - LaCour even fabricated the research grants that he supposedly used to fund the surveys

How Was This Discovered?

- Two researchers, Josh Kalla and David Broockman, attempted to replicated LaCour and Green's methods, but with the goal of reducing transphobia
- However, did not get responses rates to follow-up surveys that were similar to LaCour
- Suspicious, they investigated individual response data from LaCour (which was published along with paper)
- They found several suspicious trends in data:
  - Initial survey responses were remarkably similar to responses from another well-known paper that used same thermometer scale
  - Follow-up responses were much more highly correlated with initial responses than usually seen in literature
  - Follow-up responses seemed to be created by taking initial responses and adding positive random numbers

This Has Happened Before

- This is not the only time such fabrication has happened, unfortunately
  - One social psychology researcher in the Netherlands believed to have fabricated data in over 50 published papers
  - Not just social science: A Japanese anesthesiologist believed to have fabricated data in at least 172 papers
  - Hundreds of examples across all major research fields

Research Integrity More Broadly

- Problem is not limited to outright fabrication or falsification of data
- More subtle choices by researcher can call reproducibility of results into question
  - Choice of which data to use: throw out outliers, focus on subsample analysis, pilot several designs of experiment
  - Choice of which regressions to run
  - Choice of which statistical tests to use
  - These issues put under the general umbrella of "p-hacking"

Motivating Example

- Suppose you are running a simple experiment
  - Randomly assign people to either hot or cold room
  - Ask whether they would like \$10 now (impatient) or \$11 tomorrow (patient)
- Suppose your sample size is $N = 2$ individuals, one to each treatment
- Suppose you find that the person in the hot room takes the patient option and the person in the cold room takes the impatient option
- Can you conclude that warmer rooms cause people to act more patient?
  - No; even if temperature has no effect on patience, there is a 50% chance of getting the result we did
  - This is because there is 50% chance that we just happened to select the more patient person for the hot treatment
  - Thus in this example, the p-value is 0.5

Review of Hypothesis Testing

- More generally, are testing whether we can accept or reject a certain hypothesis
- Typically, the *null hypothesis* predicts that there will be no difference between our treatments, while the *alternate hypothesis* predicts there will be a difference
- In temperature example:
  - Null hypothesis: temperature has no effect on patience
  - Alternate hypothesis: temperature causes people to act more patient

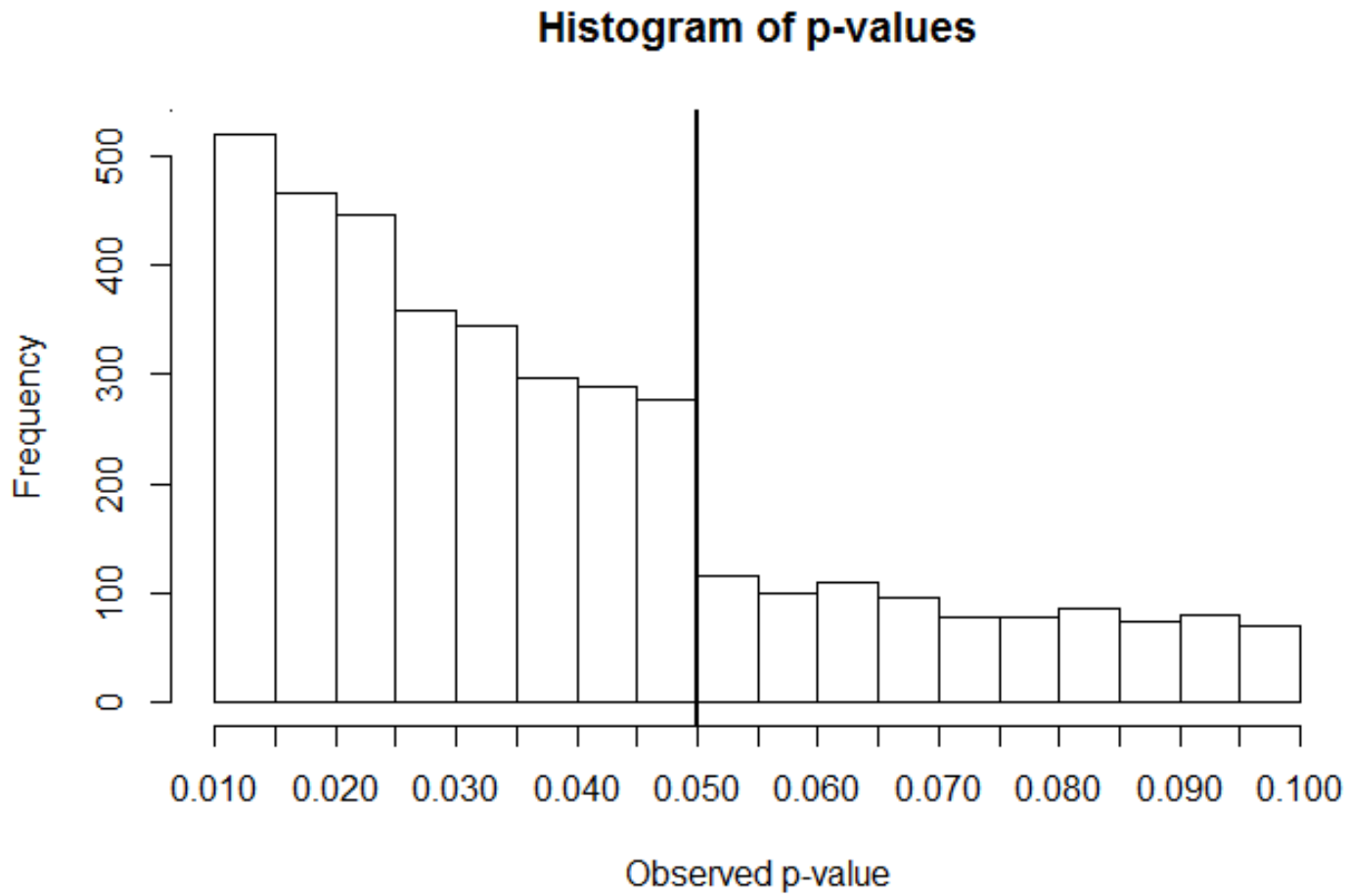
Review of *p*-values

- The *p*-value measures the probability of getting the observed result *under the null hypothesis*
  - A *p*-value close to 0 means that there is only a small likelihood that results are due to chance
  - A *p*-value close to 1 means that there is a high likelihood that results are due to chance
- For historical and largely arbitrary reasons, a *p*-value of 0.05 or less is considered "statistically significant"
- If we look at *p*-values across an entire field, distribution should be smooth

Research Integrity

- Consider all the choices we made when running the temperature experiment:
  - What temperature to make the rooms
  - What size prizes to use
- And choices made when analyzing the data:
  - Throw out responses from that one subject that fell asleep
  - Maybe we should control for gender, or GPA, or income, or ...
- If we make these choices in an attempt to get $p = 0.05$ (even subconsciously), then these are all ways of p-hacking

Visualization of p-hacking



Data: 3627 p-values reported in 3 different pyschology journals, from Masicampo and LaLande (2012)

Returning to Example

- Now suppose sample size was $N = 100$, with 50 people in each treatment
- Suppose you find that all 50 people in the hot room take the patient option and all 50 people in the cold room take the impatient option
- Now can you conclude that temperature has an effect on patience?
  - Almost certainly yes: getting this result by chance is the null was true is extremely unlikely
  - If we assume that people are equally likely to be patient or impatient under null (which might not be true), then getting this result is like flipping 50 heads in a row on a fair coin
  - Thus the p-value is essentially 0

What To Do?

1. Open up the data
   - Make all researcher publish raw data and code
   - Issue: what about proprietary/sensitive data?
2. Encourage replication
   - Don't put too much credence in results until they have been replicated independently
   - Issue: how to incentivize more replications?
3. Encourage pre-analysis plans
   - Force researchers to register experimental designs and analysis plans (eg which regressions to run) before running experiment
   - Would alleviate p-hacking and *file-drawer* effect (papers with null results not seeing the light of day)

Replication Can Work

- Recall that Broockman and Kalla were attempting to replicate LaCour and Green's canvassing methods to reduce transphobia
- Replication paper was recently published in Science (same journal that publish now-discredited LaCour and Green paper)
- Data: 1825 voters in Florida
- What they found:
  - Both transgender and non-transgender canvassers effective at changing opinions
  - These changes lasted at least 3 months
  - Key seems to be forcing respondents to do "perspective-taking" rather than logical or legal arguments